

DEVICE AND METHOD FOR EXTRACTING TITLE AREA FROM DOCUMENT IMAGE AND DOCUMENT RETRIEVING METHOD

Publication number: JP2000148788

Publication date: 2000-05-30

Inventor: OUCHI SHIGEKI

Applicant: RICOH KK

Classification:

- international: G06K9/20; G06F17/21; G06F17/27; G06F17/30;
G06K9/20; G06K9/20; G06F17/21; G06F17/27;
G06F17/30; G06K9/20; (IPC1-7): G06F17/30;
G06F17/27; G06K9/20

- European:

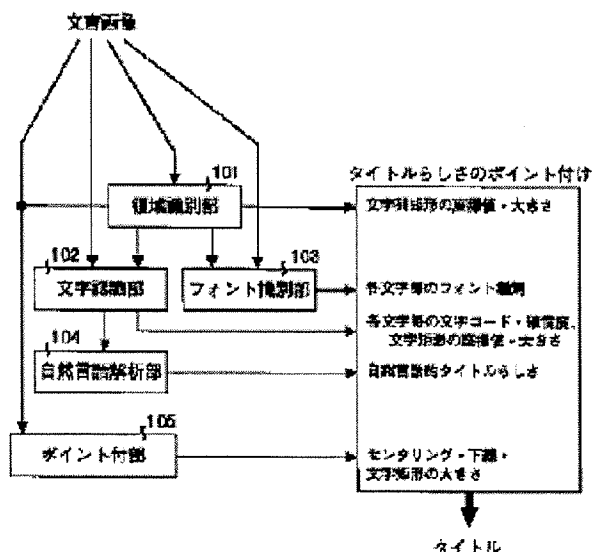
Application number: JP19980328806 19981105

Priority number(s): JP19980328806 19981105

Report a data error here

Abstract of JP2000148788

PROBLEM TO BE SOLVED: To improve the accuracy of title extraction and the convenience of document retrieval by analyzing the likelihood of a tile which is linguistically natural according to a character code obtained as a recognition result and giving points to the title likelihood by using centering, underscoring, the size of a character rectangle, etc. **SOLUTION:** A character recognition part 102 obtains the character code, accuracy, coordinate values of the character rectangle, and size of each character as points as title likelihood. A font identification part 103 obtains the font kind of each character as points as the title likelihood. The character code is supplied to a natural language analyzing routine of a natural language analysis part 104 to give points as natural language title likelihood and the points of title likelihood are given to a character string area having a word ending matching the word ending appearing frequency in a title through a natural language process. Further, total points of title likelihood are calculated by using a centering process, an underscoring process, the size of the character string rectangle, etc., to identify the title.



Family list**4** family members for: **JP2000148788**

Derived from 3 applications

[Back to JP2000148](#)

- 1 DEVICE AND METHOD FOR EXTRACTING TITLE AREA FROM DOCUMENT IMAGE AND DOCUMENT RETRIEVING METHOD**
Inventor: OUCHI SHIGEKI **Applicant:** RICOH KK
EC: **IPC:** *G06K9/20; G06F17/21; G06F17/27* (+10)
Publication info: **JP2000148788 A** - 2000-05-30
- 2 INFORMATION RETRIEVAL METHOD AND DEVICE AND RECORDING MEDIUM**
Inventor: OUCHI SHIGEKI **Applicant:** RICOH KK
EC: **IPC:** *G06K9/03; G06F17/30; G06T1/00* (+9)
Publication info: **JP2000259847 A** - 2000-09-22
- 3 METHOD AND SYSTEM FOR EXTRACTING TITLE FROM DOCUMENT IMAGE**
Inventor: OUCHI SHIGEKI (JP) **Applicant:**
EC: G06K9/20L4 **IPC:** *G06K9/20; G06K9/20*; (IPC1-7): G06K9/34 (+1)
Publication info: **US7099507 B2** - 2006-08-29
US2003156754 A1 - 2003-08-21

Data supplied from the **esp@cenet** database - Worldwide

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2000-148788

(P2000-148788A)

(43)公開日 平成12年5月30日(2000.5.30)

(51)Int.Cl. ⁷	識別記号	F I	テーマコード(参考)
G 0 6 F 17/30		G 0 6 F 15/40	3 7 0 B 5 B 0 0 9
17/27		G 0 6 K 9/20	3 4 0 5 B 0 2 9
G 0 6 K 9/20	3 4 0	G 0 6 F 15/20	5 5 0 F 5 B 0 7 5
		15/401	3 1 0 A

審査請求 未請求 請求項の数11 F D (全 9 頁)

(21)出願番号 特願平10-328806

(22)出願日 平成10年11月5日(1998.11.5)

(71)出願人 000006747

株式会社リコー

東京都大田区中馬込1丁目3番6号

(72)発明者 大内 茂樹

東京都大田区中馬込1丁目3番6号 株式

会社リコー内

Fターム(参考) 5B009 QA12

5B029 AA01 BB02 CC27

5B075 ND03 NK02 NK04 NK32 NK39

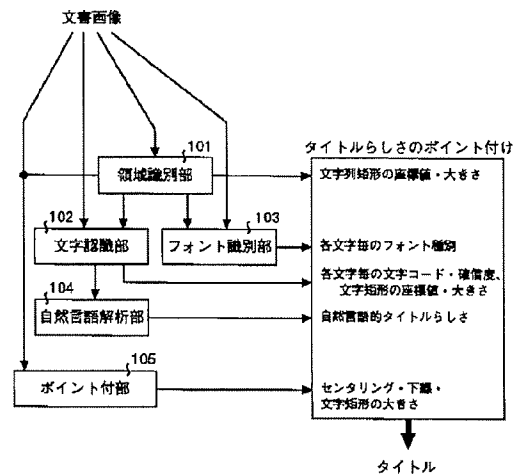
UU06

(54)【発明の名称】 文書画像からのタイトル領域抽出装置およびタイトル領域抽出方法、並びに文書検索方法

(57)【要約】

【課題】 特定の文書形式に依存せずにタイトル固有の特徴をポイントとして用いることにより、ポイント数の多い文字列領域をタイトルとして自動抽出し、タイトル抽出の的確性および文書検索時の利便性を向上させること。

【解決手段】 領域識別部101で切り出された文字列矩形に対し、該文字列矩形内の文字認識を行う文字認識部102と、上記文字列矩形に対し、該文字列矩形内の各文字毎のフォント識別を行うフォント識別部103と、文字認識部102の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析部104と、上記文字列矩形に対し、センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付部105と、を備えた。



【特許請求の範囲】

【請求項1】 画像入力装置から入力された文書画像から文字列領域を矩形で切り出す領域識別手段を有し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出装置において、前記領域識別手段で切り出された文字列矩形に対し、該文字列矩形内の文字認識を行う文字認識手段と、前記領域識別手段で切り出された文字列矩形に対し、該文字列矩形内の各文字毎のフォント識別を行うフォント識別手段と、前記文字認識手段の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段と、前記領域識別手段で切り出された文字列矩形に対し、センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段と、を備えたことを特徴とする文書画像からのタイトル領域抽出装置。

【請求項2】 入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域内の文字コードを認識し、文字コード識別の確信度が一定のしきい値以上であるかを判断する第1の工程と、前記第1の工程で一定のしきい値以上である場合、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第2の工程と、を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項3】 入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域内の文字認識を実行し、該文字認識時に文字列矩形内の文字数を求める第1の工程と、文書のタイトルの文字数を用い、前記文字数と比較し、文字矩形数が所定値内であるかを判断する第2の工程と、前記第2の工程で、文字矩形数が所定値内である場合に、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項4】 入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第1の工程と、前記第1の工程の結果、体言止めになっている領域であるかを判断する第2の工程と、前記第2の工程で体言止めになっている領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むことを特徴とする文書画像からのタイトル領域抽出方

法。

【請求項5】 入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第1の工程と、前記第1の工程の結果、タイトルに頻出する語尾の統計情報辞書と前記文字列領域内の文字コード列とを比較し、高頻出度の語尾と一致するものを語尾に含む文字列領域であるかを判断する第2の工程と、前記第2の工程の領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項6】 入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域に対してフォント識別処理を実行する第1の工程と、前記フォント識別処理の結果に基づいて、文字のフォントスタイルを判別し、特定のフォントを用いている文字領域であるかを判断する第2の工程と、前記第2の工程で特定のフォントを用いている文字領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項7】 入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域に対してフォント識別処理を実行する第1の工程と、前記フォント識別処理の結果に基づいて、フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき、出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断する第2の工程と、前記第2の工程で判断した文字領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項8】 入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域内の各文字矩形のアスペクト比を求める第1の工程と、前記アスペクト比に基づいて倍角文字であるかを判断する第2の工程と、前記倍角文字であると判断した文字領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項9】 入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域に対して文字認識処理を実行する第1の工程と、前記文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求める第2の工程と、前記合計値が前記文字矩形領域のほぼ半分であるかを判断する第3の工程と、前記第3の工程でほぼ半分であると判定された文字列領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第4の工程と、を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項10】 前記タイトルらしさのポイント加算の可否判断に用いる基準値は、ユーザ単位の入力文書形式に合わせて学習して得られる最適値とし、可変・設定されることを特徴とする請求項2ないし9の何れか一つに記載の文書画像からのタイトル領域抽出方法。

【請求項11】 文書画像を認識し、その結果に対して言語処理を行ってキーワードを抽出する第1の工程と、前記第1の工程で抽出されたキーワードと、請求項2ないし10の何れか一つに記載の文書画像からのタイトル領域抽出方法に基づいて抽出したタイトルとを併記する第2の工程と、前記第2の工程で併記されたタイトルを用いて文書検索を実行する第3の工程と、を含むことを特徴とする文書検索方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ファクシミリやイメージスキャナ等の画像入力装置から入力された文書画像データのデータベースから、検索の利便性を向上させるために、文書内容を的確に表現するような文書中の領域をタイトル領域として抽出する文書画像からのタイトル領域抽出装置およびタイトル領域抽出方法、並びに文書検索方法に関する。

【0002】

【従来の技術】従来、文書画像を検索する際には、後の検索時の利便性を図るために、画像入力装置からの文書画像の入力とは別にオペレータが手作業で、その文書の内容を的確に表現するタイトル情報やキーワード情報を抽出／作成して付加したり、定形文書に対しては、文書中の特定の位置（文字列）をタイトル・キーワードとして切り出していた。

【0003】また、非定形文書に対してレイアウト的特徴のみを用いてタイトルを抽出する参考技術文献が、例えば、特開平9-134406号公報の『文書画像からのタイトル抽出装置および方法』、特開平5-274471号公報の『イメージ文書のタイトル領域抽出処理方法』が開示されている。

【0004】

【発明が解決しようとする課題】しかしながら、上記に示されるような従来の技術にあっては、オペレータによるタイトル情報やキーワード情報の付加は文書量が多くなるにしたがって作業量も増加するため、作業負担の増大化を招来させてしまう。また、特定の位置の自動切り出しを行うと、定形文書のみを対象とするので、非定形文書には利用することができず、利便性に欠けるといった問題点があった。

【0005】従来より開示されている特開平9-134406号公報・特開平5-274471号公報にあっては、レイアウト的特徴にのみ注目してタイトル抽出を行っているため、文書内容を的確に表現するタイトルの的中率が必ずしも満足できるものではなく、後の文書検索等に支障をきたす等の問題点があった。

【0006】本発明は、上記に鑑みてなされたものであって、特定の文書形式に依存せずにタイトル固有の特徴をポイントとして用いることにより、ポイント数の多い文字列領域をタイトルとして自動抽出し、タイトル抽出の的確性および文書検索時の利便性を向上させることを目的とする。

【0007】

【課題を解決するための手段】上記の目的を達成するために、請求項1に係る文書画像からのタイトル領域抽出装置にあっては、画像入力装置から入力された文書画像から文字列領域を矩形で切り出す領域識別手段を有し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出装置において、前記領域識別手段で切り出された文字列矩形に対し、該文字列矩形内の文字認識を行う文字認識手段と、前記領域識別手段で切り出された文字列矩形に対し、該文字列矩形内の各文字毎のフォント識別を行うフォント識別手段と、前記文字認識手段の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段と、前記領域識別手段で切り出された文字列矩形に対し、センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段と、を備えたものである。

【0008】また、請求項2に係る文書画像からのタイトル領域抽出方法にあっては、入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域内の文字コードを認識し、文字コード識別の確信度が一定のしきい値以上であるか否かを判断する第1の工程と、前記第1の工程で一定のしきい値以上である場合、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第2の工程と、を含むものである。

【0009】また、請求項3に係る文書画像からのタイ

トル領域抽出方法にあつては、入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域内の文字認識を実行し、該文字認識時に文字列矩形内の文字数を求める第1の工程と、文書のタイトルの文字数を用い、前記文字数と比較し、文字矩形数が所定値内であるか否かを判断する第2の工程と、前記第2の工程で、文字矩形数が所定値内である場合に、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むものである。

【0010】また、請求項4に係る文書画像からのタイトル領域抽出方法にあつては、入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第1の工程と、前記第1の工程の結果、体言止めになっている領域であるかを判断する第2の工程と、前記第2の工程で体言止めになっている領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むものである。

【0011】また、請求項5に係る文書画像からのタイトル領域抽出方法にあつては、入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第1の工程と、前記第1の工程の結果、タイトルに頻出する語尾の統計情報辞書と前記文字列領域内の文字コード列とを比較し、高頻出度の語尾と一致するものを語尾に含む文字列領域であるかを判断する第2の工程と、前記第2の工程の領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むものである。

【0012】また、請求項6に係る文書画像からのタイトル領域抽出方法にあつては、入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域に対してフォント識別処理を実行する第1の工程と、前記フォント識別処理の結果に基づいて、文字のフォントスタイルを判別し、特定のフォントを用いている文字領域であるかを判断する第2の工程と、前記第2の工程で特定のフォントを用いている文字領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工

程と、を含むものである。

【0013】また、請求項7に係る文書画像からのタイトル領域抽出方法にあつては、入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列領域に対してフォント識別処理を実行する第1の工程と、前記フォント識別処理の結果に基づいて、フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき、出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断する第2の工程と、前記第2の工程で判断した文字領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むものである。

【0014】また、請求項8に係る文書画像からのタイトル領域抽出方法にあつては、入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列矩形内の各文字矩形のアスペクト比を求める第1の工程と、前記アスペクト比に基づいて倍角文字であるかを判断する第2の工程と、前記倍角文字であると判断した文字領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第3の工程と、を含むものである。

【0015】また、請求項9に係る文書画像からのタイトル領域抽出方法にあつては、入力された文書画像から文字列領域を矩形で切り出し、前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する文書画像からのタイトル領域抽出方法において、前記文字列矩形に対して文字認識処理を実行する第1の工程と、前記文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求める第2の工程と、前記合計値が前記文字矩形領域のほぼ半分であるかを判断する第3の工程と、前記第3の工程でほぼ半分であると判定された文字列領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する第4の工程と、を含むものである。

【0016】また、請求項10に係る文書画像からのタイトル領域抽出方法にあつては、前記タイトルらしさのポイント加算の可否判断に用いる基準値は、ユーザ単位の入力文書形式に合わせて学習して得られる最適値とし、可変・設定されるものである。

【0017】また、請求項11に係る文書検索方法にあつては、文書画像を認識し、その結果に対して言語処理を行ってキーワードを抽出する第1の工程と、前記第1の工程で抽出されたキーワードと、請求項2ないし10の何れか一つに記載の文書画像からのタイトル領域抽出

方法に基づいて抽出したタイトルとを併記する第2の工程と、前記第2の工程で併記されたタイトルを用いて文書検索を実行する第3の工程と、を含むものである。

【0018】

【発明の実施の形態】以下、本発明の文書画像からのタイトル領域抽出装置およびタイトル領域抽出方法、並びに文書検索方法について添付図面を参照して説明する。

【0019】図1は、本発明の実施の形態に係るタイトル領域抽出処理を行うシステム構成を示すブロック図である。図において、101はファクシミリやイメージスキャナ等の画像入力装置（図示せず）から入力された文書画像から文字列領域を矩形で切り出す領域識別手段としての領域識別部、102は領域識別部101の識別結果に基づいて文字認識を行う文字認識手段としての文字認識部、103は領域識別部101の識別結果に基づいてフォント識別を行うフォント識別手段としてのフォント識別部、104は文字認識部102の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段としての自然言語解析部、105は従来から用いられているセンタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段としてのポイント付部である。

【0020】図1の構成において、画像入力装置（図示せず）から文書画像が入力されると、スキュー補正等の前処理を行い、領域識別部101により領域識別処理を実行し、文字列矩形の座標値・大きさの情報を得る。次いで、領域識別部101による領域識別処理の結果を用い、文字認識部102による文字認識、およびフォント識別部103によるフォント識別を行う。

【0021】文字認識部102では各文字毎の文字コード・確信度、文字矩形の座標値・大きさがタイトルらしさのポイント付けとして得られる。また、フォント識別部103では各文字毎のフォント種別がタイトルらしさのポイント付けとして得られる。

【0022】また、文字認識部102により得られる文字コードは、自然言語解析部104自然言語解析ルーチンにも供給され、自然言語的タイトルらしさ、つまり、体言止めになっている領域のタイトルらしさのポイントを与える。さらに、自然言語処理において、タイトルに類出する語尾の統計情報辞書と文字領域内の文字コード列とを比較し、高頻出度の語尾と一致するものを語尾に含む文字列領域にタイトルらしさのポイントを与える。

【0023】また、上述の各ポイントらしさのポイントに加え、従来から用いられているセンタリング処理・下線処理・文字列矩形の大きさ等も用いてタイトルらしさの合計ポイント进行計算し、タイトルを識別する。

【0024】次に、図3～図8に示すフローチャートを参照し、本発明の一連のタイトル抽出処理方法について順に説明する。なお、このタイトル抽出処理方法は、図1の構成によって複数の組み合わせあるいは単独、あるいは

は選択的に行ってことができる。

【0025】図3は、実施の形態に係る第1のタイトル抽出方法を示すフローチャートであり、文字コード識別の確信度が一定のしきい値以上であった場合にタイトルらしさのポイントを加算する例について示している。まず、文書入力装置（図示せず）から文書画像を入力し（S301）、領域識別部101により文字列領域を識別する（S302）。続いて、上記文字列領域内の文字コードを認識し、文字コード識別の確信度が一定のしきい値以上であるか否かを判断する（S303）。ここで、一定のしきい値以上であると判断した場合、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する（S304）。

【0026】図4は、実施の形態に係る第2のタイトル抽出方法を示すフローチャートである。まず、文書入力装置（図示せず）から文書画像を入力し（S401）、領域識別部101により文字列領域を識別する（S402）。続いて、文字認識時に文字列矩形内の文字数を求める（S403）。そして、文書のタイトルの文字数を用い、上記文字数と比較し（S404）、文字矩形数が所定値内であるか否かを判断する（S405）。ここで、文字矩形数が所定値内であると判断すると、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する（S406）。

【0027】すなわち、文字列領域内の文字コード認識時に文字列矩形内の文字数を求め、別途辞書情報として文書のタイトルの文字数の統計を用いて比較し、タイトルらしい文字数の文字列矩形に対してタイトルらしさのポイントを与える。

【0028】図5は、実施の形態に係る第3のタイトル抽出方法を示すフローチャートである。まず、文書入力装置（図示せず）から文書画像を入力し（S501）、領域識別部101により文字列領域を識別する（S502）。続いて、文字列領域内の文字コードの認識結果に対して自然言語処理を行い（S503）、所定事項の領域、例えば、体言止めになっている領域か否かを判断する（S504）。ここで、所定事項の領域であると判断すると、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する（S505）。

【0029】また、上述の言語処理において、タイトルに類出する語尾の統計情報辞書と文字領域内の文字コード列とを比較し、高頻出度の語尾と一致するものを語尾に含む文字列領域にタイトルらしさのポイントを与えてもよい。

【0030】図6は、実施の形態に係る第4のタイトル抽出方法を示すフローチャートである。まず、文書入力装置（図示せず）から文書画像を入力し（S601）、領域識別部101により文字列領域を識別する（S602）。続いて、フォント識別処理を行い（S603）、所定のフォント（フォントスタイル）を含む領域である

か否かを判断する（S604）。つまり、文字のフォントスタイルを判別し、特定のフォントを用いている文字領域であるか、あるいは、フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき、出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断する。そして、これらの領域であると判断した場合に、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する（S605）。

【0031】図7は、実施の形態に係る第5のタイトル抽出方法を示すフローチャートである。まず、文書入力装置（図示せず）から文書画像を入力し（S701）、領域識別部101により文字列領域を識別する（S702）。続いて、文字列矩形内の各文字矩形のアスペクト比を求め（S703）、アスペクト比が横：縦＝2：1に近い値となる文字矩形が文字列矩形内の文字矩形数中の一定の割合以上を占めているか否かを判断する（S704）。ここで、一定以上の割合を占めていれば、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する（S705）。

【0032】図8は、実施の形態に係る第6のタイトル抽出方法を示すフローチャートである。まず、文書入力装置（図示せず）から文書画像を入力し（S801）、領域識別部101により文字列領域を識別する（S802）。続いて、文字認識処理を行い（S803）、文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求める（S804）。そして、その合計値が文字矩形領域のほぼ半分であるか否かを判断する（S805）。ここで、合計値が文字矩形領域のほぼ半分であれば、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出する（S806）。

【0033】ところで、上述した実施の形態において必要となるしきい値を固定せずに、各ユーザの入力する文書の対応させて学習し、各ユーザの使用する文書形式に対して最適なしきい値を可変的に求め、初期値から変更・設定してもよい。

【0034】さらに、上述の如く求められる一時的なポイントを基に、図2に示すように二次的な組み合わせにより、倍角文字や均等割付けの判定を行い、それらに対してタイトルらしさのポイントを与えることも可能である。

【0035】これを付言すると、文字コードの認識時に得られる文字列矩形内の各文字矩形領域の大きさを用い、文字矩形領域のアスペクト比を算出することによって倍角文字を判定し、該倍角文字を用いている文字列領域に対してタイトルらしさのポイントを与える。

【0036】また、文字矩形領域とそれが属する文字列領域の大きさと、文字コードの認識時に得られる文字列矩形内の文字数とを用いて矩形内の文字密度を算出し、

その値によって均等割付け判定を行う。そして、均等割付けされたと判定された文字列領域に対してタイトルらしさのポイントを与える。

【0037】ところで、上述したタイトル領域抽出方法を用いて情報検索を行うことも実現可能である。図9は、実施の形態に係る情報検索方法を示すフローチャートである。まず、文書画像を認識し（S901）、その結果に対して言語処理を行ってキーワードを抽出する（S902）。さらに、上記抽出されたキーワードと、前述のタイトル領域抽出方法によって抽出したタイトルとを併記し（S903）、その併記タイトルを用いて文書検索を実行する（S904）。これにより、検索時における利便性が向上する。

【0038】

【発明の効果】以上説明したように、本発明に係る文書画像からのタイトル領域抽出装置（請求項1）によれば、入力された文書画像から文字列領域を矩形で切り出す領域識別手段を有し、その文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し、タイトルを抽出する際に、領域識別手段で切り出された文字列矩形に対し、該文字列矩形内の文字認識を行う文字認識手段と、領域識別手段で切り出された文字列矩形に対し、該文字列矩形内の各文字毎のフォント識別を行うフォント識別手段と、文字認識手段の認識結果で得られる文字コードに基づいて自然言語のタイトルらしさを解析する自然言語解析手段と、領域識別手段で切り出された文字列矩形に対し、センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段とを設け、特定の文書形式に依存せずにタイトル固有の特徴をポイント付けとして用いるため、ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し、かつタイトル抽出の的確性および文書検索時の利便性を向上させる装置を提供することができる。

【0039】また、本発明に係る文書画像からのタイトル領域抽出方法（請求項2）によれば、文字列領域内の文字コードを認識し、文字コード識別の確信度が一定のしきい値以上であるかを判断し、一定のしきい値以上である場合、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出するため、ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し、かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【0040】また、本発明に係る文書画像からのタイトル領域抽出方法（請求項3）によれば、文字列領域内の文字認識を実行し、該文字認識時に文字列矩形内の文字数を求め、文書のタイトルの文字数を用い、上記文字数と比較し、文字矩形数が所定値内であるかを判断し、文字矩形数が所定値内である場合、該当する文字列領域にタイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出するため、ポイント数の多い

文字列領域をタイトルとして自動抽出することが実現し、かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【0041】また、本発明に係る文書画像からのタイトル領域抽出方法（請求項4）によれば、文字列領域内の文字コードの認識結果に対して自然言語処理を実行し、その結果、体言止めになっている領域であるかを判断し、体言止めになっている領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出するため、ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し、かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【0042】また、本発明に係る文書画像からのタイトル領域抽出方法（請求項5）によれば、文字列領域内の文字コードの認識結果に対して自然言語処理を実行し、その結果、タイトルに頻出する語尾の統計情報辞書と文字列領域内の文字コード列とを比較し、高頻出度の語尾と一致するものを語尾に含む文字列領域であるかを判断し、その領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出するため、ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し、かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【0043】また、本発明に係る文書画像からのタイトル領域抽出方法（請求項6）によれば、文字列領域に対してフォント識別処理を実行し、その結果に基づいて、文字のフォントスタイルを判別し、特定のフォントを用いている文字領域であるかを判断し、特定のフォントを用いている文字領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出するため、ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し、かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【0044】また、本発明に係る文書画像からのタイトル領域抽出方法（請求項7）によれば、文字列領域に対してフォント識別処理を実行し、その結果に基づいて、フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき、出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断し、該判断した文字領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出するため、ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し、かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【0045】また、本発明に係る文書画像からのタイトル領域抽出方法（請求項8）によれば、文字列矩形内の各文字矩形のアスペクト比を求め、そのアスペクト比に

基づいて倍角文字であるかを判断し、倍角文字であると判断した文字領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出するため、ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し、かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【0046】また、本発明に係る文書画像からのタイトル領域抽出方法（請求項9）によれば、文字列矩形に対して文字認識処理を実行し、文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求め、その合計値が文字矩形領域のほぼ半分であるかを判断し、ほぼ半分であると判定された文字列領域に対し、タイトルらしさのポイントを加算し、その合計値によりタイトル領域を決定・抽出するため、ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し、かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【0047】また、本発明に係る文書画像からのタイトル領域抽出方法（請求項10）によれば、請求項2ないし9の何れか一つに記載の文書画像からのタイトル領域抽出方法において、タイトルらしさのポイント加算の可否判断に用いる基準値を、ユーザ単位の入力文書形式に合わせて学習して得られる最適値を用いて可変・設定することにより、よりの確なタイトルの自動抽出が実現する。

【0048】また、本発明に係る文書検索方法（請求項11）によれば、文書画像を文字認識し、その結果に対して言語処理を行って抽出されたキーワードと、請求項2ないし9の何れか一つに記載の文書画像からのタイトル領域抽出方法に基づいて抽出したタイトルとを併記し、該併記されたタイトル、すなわち、よりの確なタイトルを用いて文書検索を実行するため、文書検索時における利便性が向上する。

【図面の簡単な説明】

【図1】本発明の実施の形態に係るタイトル領域抽出処理を行うシステム構成を示すブロック図である。

【図2】本発明の実施の形態に係るタイトル領域抽出処理に用いられるタイトルらしさのポイントうち、二次的に求められるものを示すブロック図である。

【図3】本発明の実施の形態に係る第1のタイトル抽出方法を示すフローチャートである。

【図4】本発明の実施の形態に係る第2のタイトル抽出方法を示すフローチャートである。

【図5】本発明の実施の形態に係る第3のタイトル抽出方法を示すフローチャートである。

【図6】本発明の実施の形態に係る第4のタイトル抽出方法を示すフローチャートである。

【図7】本発明の実施の形態に係る第5のタイトル抽出方法を示すフローチャートである。

【図8】本発明の実施の形態に係る第6のタイトル抽出

方法を示すフローチャートである。

【図9】本発明の実施の形態に係る情報検索方法を示すフローチャートである。

【符号の説明】

101 領域識別部

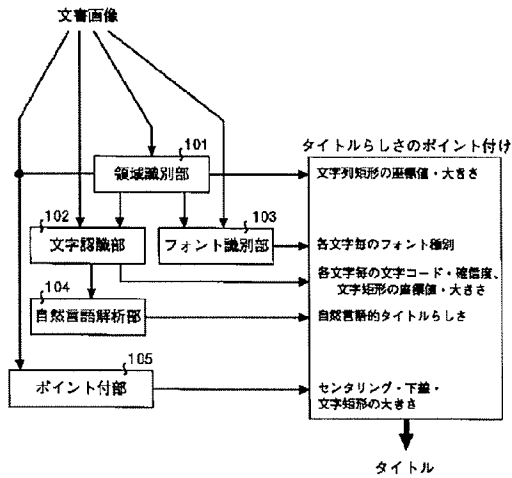
102 文字認識部

103 フォント識別部

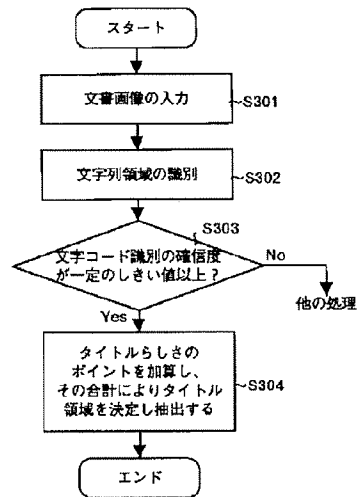
104 自然言語解析部

105 ポイント付部

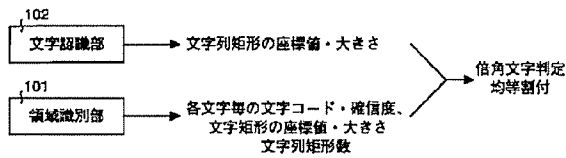
【図1】



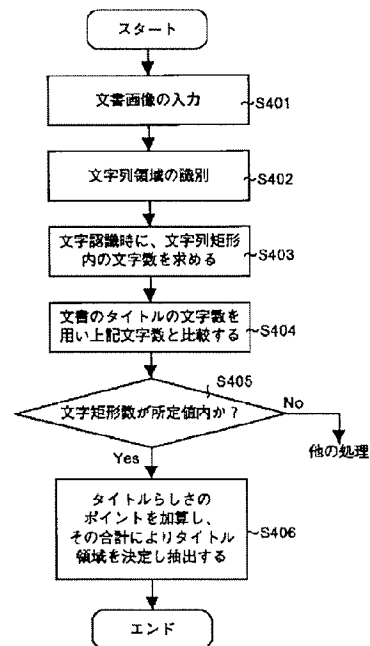
【図3】



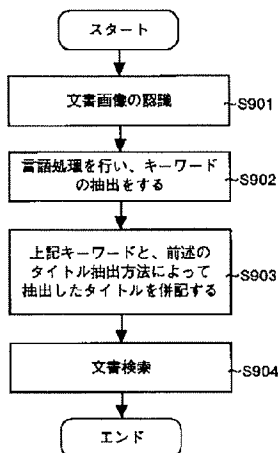
【図2】



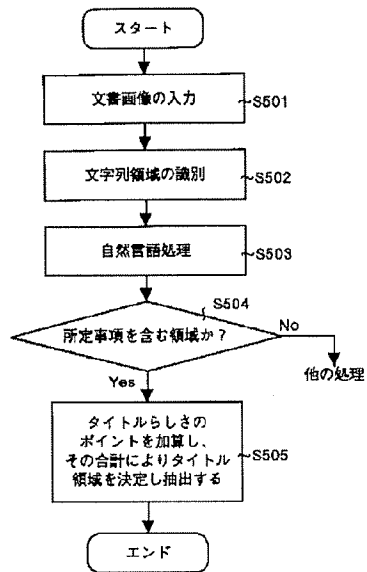
【図4】



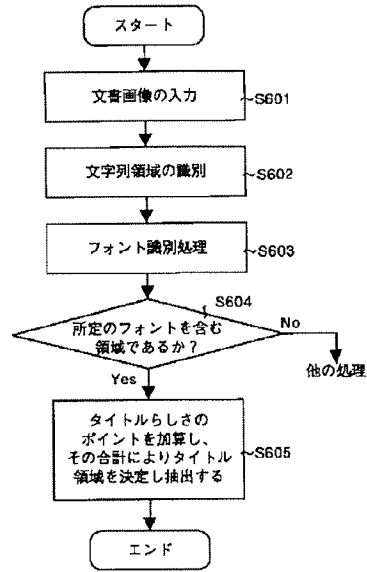
【図9】



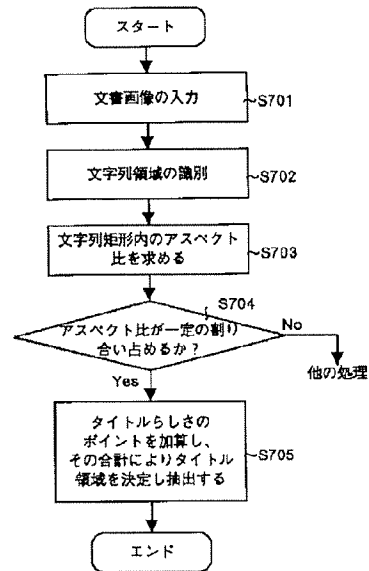
【図5】



【図6】



【図7】



【図8】

